DATA CLASSIFICATION

- **Qualitative**
  Qualitative properties are properties that are observed and can generally not be measured with a numerical result.
  1. **Nominal**
     Data with no inherent order or ranking such as gender
  2. **Ordinal**
     Data with an order series or such as sentiment options in supermarkets
- **Quantitative**
  1. **Discrete**
     Categorical data finite number of possible values of students in a class
  2. **continuous**
     Data can hold infinite number of possible values (weight of a person)

PROBABILITY VS STATISTICS

1. **Probability** Measure Of How Likely An Event Will Occur or likelyhood of an event occurrence
2. **Statistics** is a branch of mathematics that concerns the collection, organization, displaying, analysis, interpretation and presentation of data

PROBABILITY

Random experiment:- a random experiment is a **process** by which we observe something uncertain
Sample space:- Sample space of an experiment or random trial is the set of **all possible outcomes** or results of that experiment.
Event:- an event is a **one or set of outcomes** of an experiment

- **Dis-joint Event:-**doesn't have any common outcomes
- **Non Dis-joint Event:**-have some common outcomes

PROBABILITY DISTRIBUTION

PROBABILITY MASS FUNCTION:-In probability and statistics, a probability mass function (PMF) is a function that gives the **probability that a discrete random variable** is exactly equal to some value.
PROBABILITY DENSITY FUNCTION:-(RAINFALL FOR THE NEXT DAY WILL BE A RANGE 1CM-2CM)
The PDF is used to specify the probability of the random variable falling *within a particular range of values*, as opposed to taking on any one value(single value).

CUMULATIVE DENSITY FUNCTION (CDF)

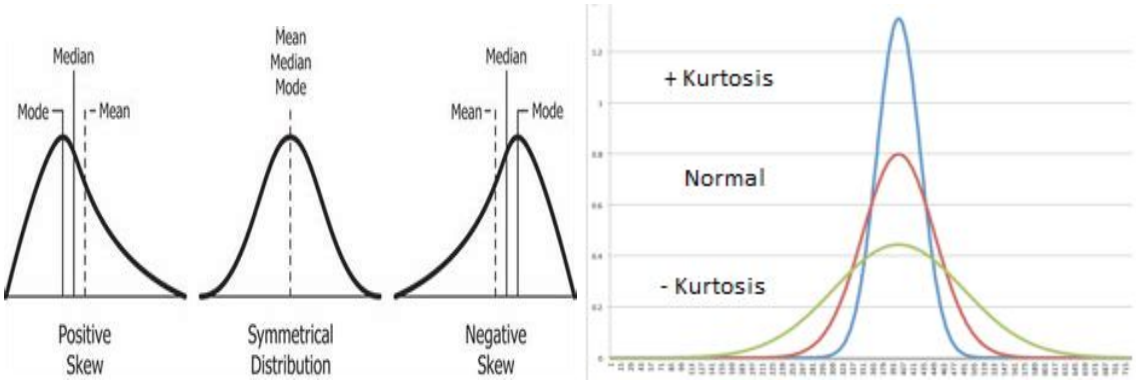
A cumulative density function (cdf) tells us the probability that a random variable takes on a **value less than or equal to x.**

Furthermore, the area under the curve of a pdf between negative infinity and x is equal to the value of x on the cdf.

NORMAL DISTRIBUTION:-

Normal distribution is a probability distribution that associates the normal random variable(it has mean at 0 and the variance equal to 1) x with a **cumulative probability**

SKEWNESS & KURTOSIS



CENTRAL LIMIT THEOREM:-

In the study of probability theory, the **central limit theorem (CLT)** states that the **distribution of sample approximates a normal distribution (also known as a "bell curve") as the sample size becomes larger,** assuming that all samples are identical in size, and regardless of the population distribution shape.

"MEAN IS EQUAL TO THE SAMPLES TO THE REAL DISTRIBUTION FOR LARGE DATASET"

TYPE OF PROBABILITY

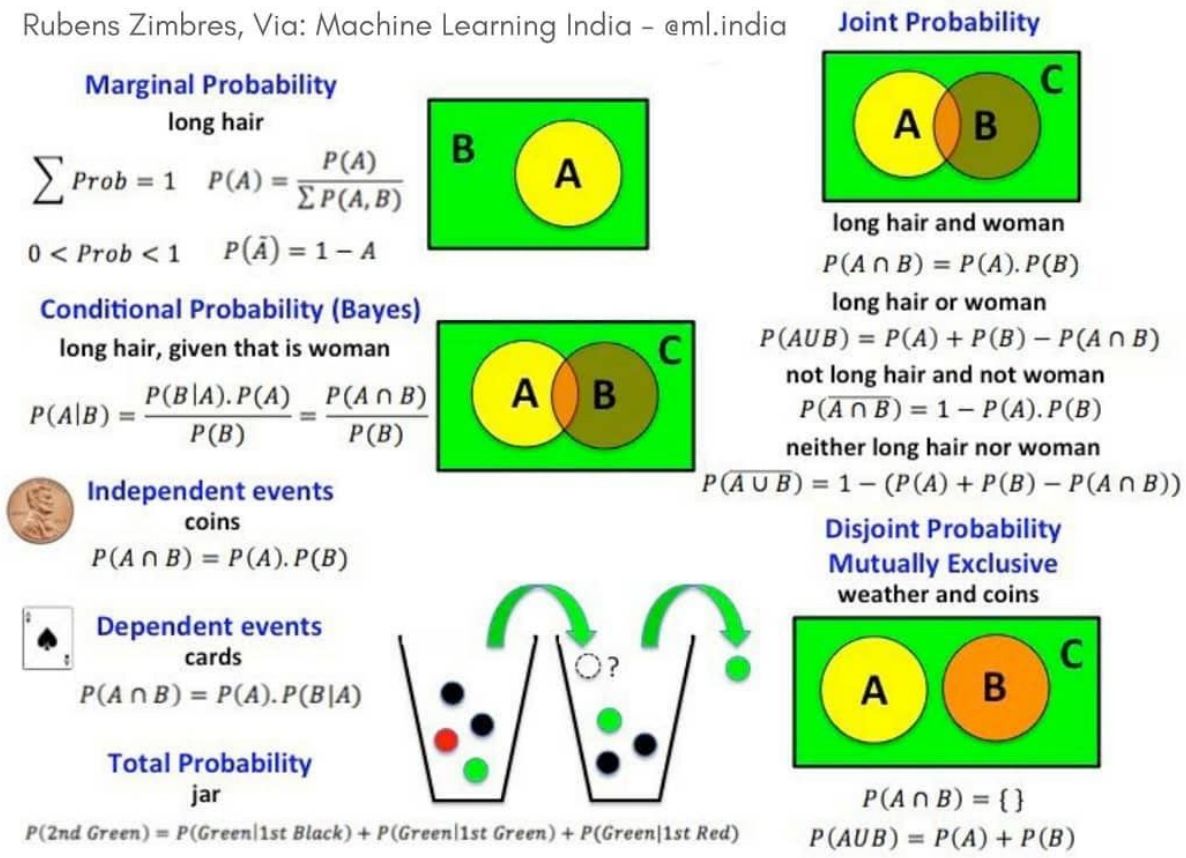
Marginal Probability:- probability of an **single event occur**

Join Probability:-A joint probability, in probability theory, refers to the probability that **two events will both occur at the same time**. In other words, joint probability is the likelihood of two events occurring together

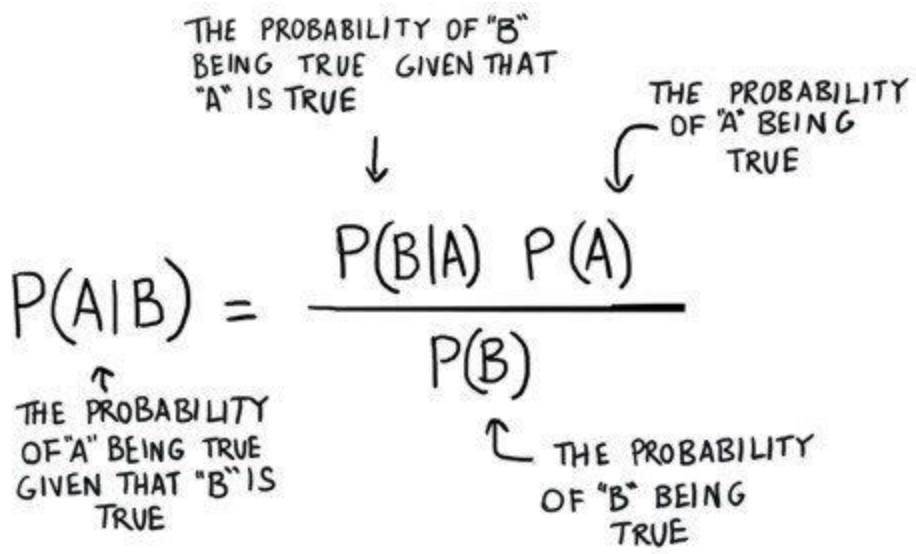
Conditional Probability:-In probability theory, conditional probability is a measure of the probability of an **event occurring, given that another event (by assumption, presumption, assertion or evidence) has already occurred.**

Probability Fundamentals

Probability is the bedrock of machine learning. You cannot develop a deep understanding and application of machine learning without it.



NAIVE BAYES THEOREM:-shows the relation between one conditional probability and its inverse



STATISTICS

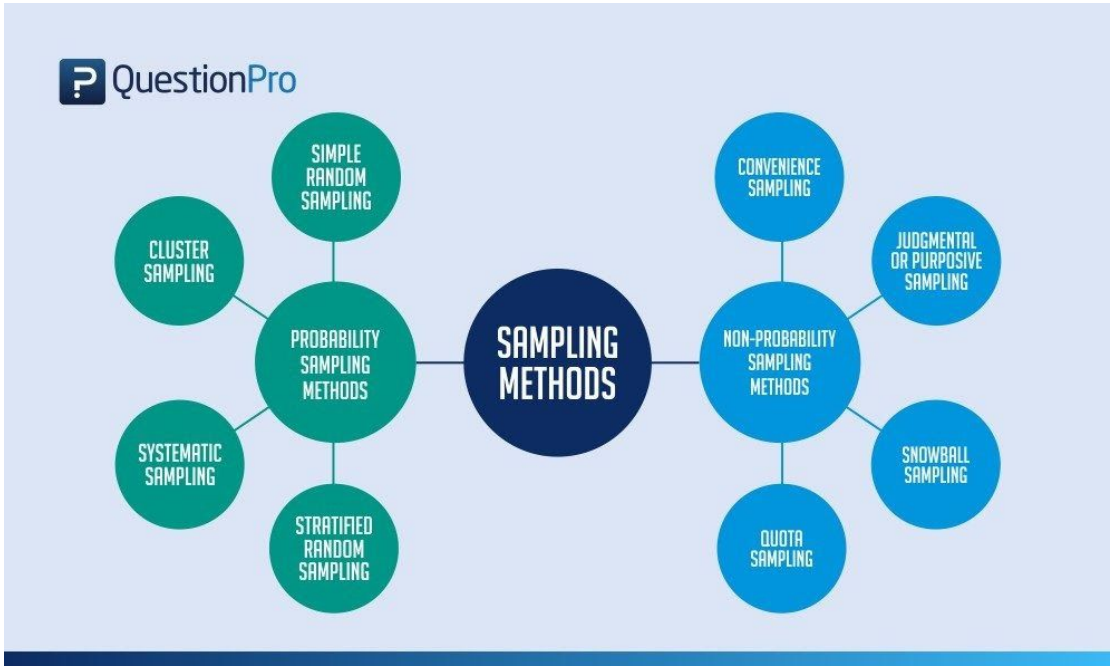
statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied.

POPULATION:-

A statistical population is any group of individuals who are the subject of a study, meaning that almost anything can make up a population so long as the individuals can be grouped together by a common feature, or sometimes two common features

SAMPLE:-

A sample is a smaller group of members of a population selected to represent the population. In order to use statistics to learn things about the population



PROBABILITY SAMPLING:-

Probability sampling is defined as a sampling technique in which the researcher chooses samples from a larger population using a method based on the theory of probability.

- **RANDOM SAMPLING**
  as the name suggests, is an e**ntirely random method** of selecting the sample.
- **SYSTEMATIC SAMPLING:- NTH RECORD**
  is when you choose every "nth" individual to be a part of the sample. For example, you can select every 5th person to be in the sample.
- **STRATIFIED SAMPLING:-**
  involves a method where the researcher divides a more extensive population into smaller groups **stratum** that usually don't overlap but represent the entire population. While sampling, organize these groups and then draw a **sample from each group separately**.
- **RANDOM CLUSTER SAMPLING** is a way to select participants randomly that are spread out geographically.

TYPES OF STATISTICS

- **DESCRIPTIVE STATISTICS**

A descriptive statistic is a summary statistic that quantitatively describes or summarizes features from a collection of information, while descriptive statistics is the process of using and analysing those statistics aims to *describe* a chunk of raw data using summary statistics, graphs, and tables.

EX:- MIN,MAX,AVG t-shirt size

- **INFERENTIAL STATISTICS**

Inferential statistics uses a small sample of data to draw *inferences* about the larger population that the sample came from.

Ex:- large,medium,large

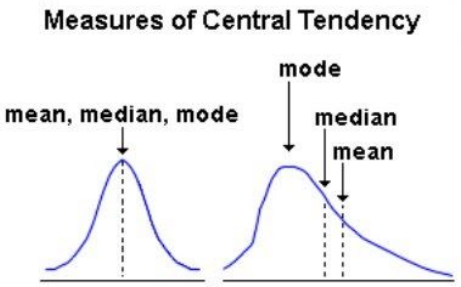
DESCRIPTIVE STATISTICS

1. **Measure of central tendency**
   - **MEAN,MEDIAN,MODE**

The **mean** is the most common measure of central tendency used by researchers and people in all kinds of professions. It is the measure of central tendency that is also referred to as the average.

The **median** is the value at the middle of a distribution of data when those data are organized from the lowest to the highest value.

The **mode** is the measure of central tendency that identifies the category or score that occurs the most frequently within the distribution of data.



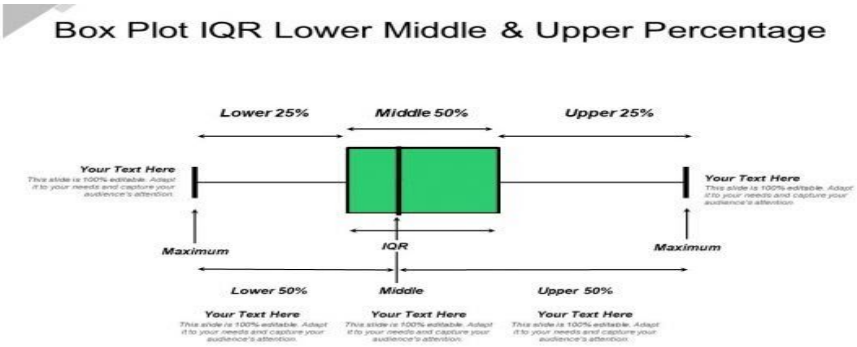
2. **Measure of variability(spread):-**

It describe the variability in a population

- The **range** of a dataset is the difference between the largest and smallest values in that dataset.
- The **interquartile range** includes the 50% of data points that fall between Q1 and Q3.

Outlier:-the data point significantly differ from others

Percentile:- It measles indicating the value below which a given percentage of observations in the given observations

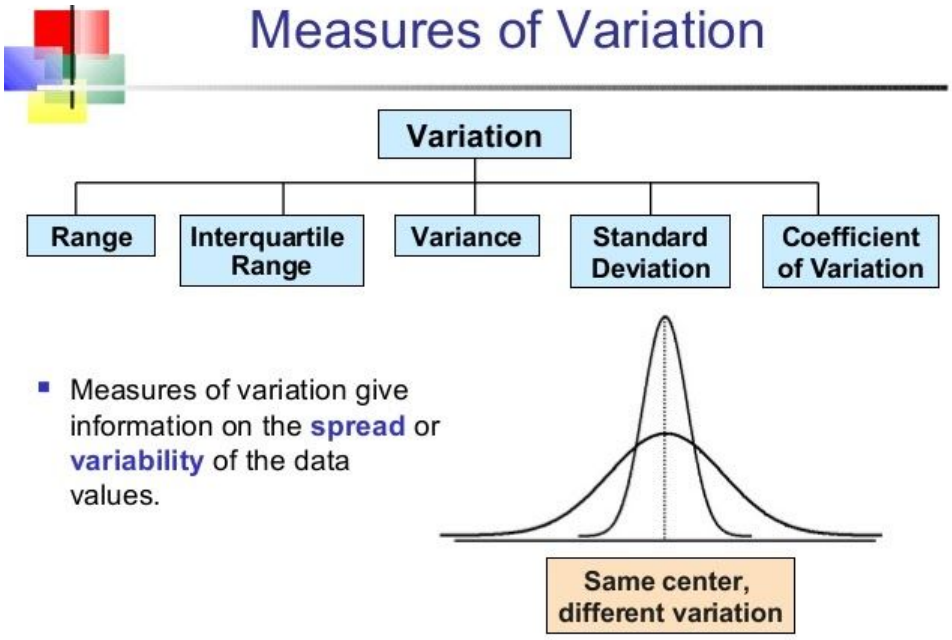


- **Variance** is the average squared difference of the values from the mean.
  1. Deviation

2. Sample Variance
3. Population Variance

- The **standard deviation** is the standard or typical difference between each data point and the mean.
- **Median absolute deviation:- (MAD)** of a data set is the average distance between each data value and the mean. Mean absolute deviation is a way to describe variation in a data set. Outlier are not that affected since we are taking modulus

$$MAD = median(|Yi - median(Yi|)$$



INFORMATION GAIN AND ENTROPY

How can we select the **root node for the decision tree??**?

ENTROPY :-(calculate on overall data 20 yes,20 no)

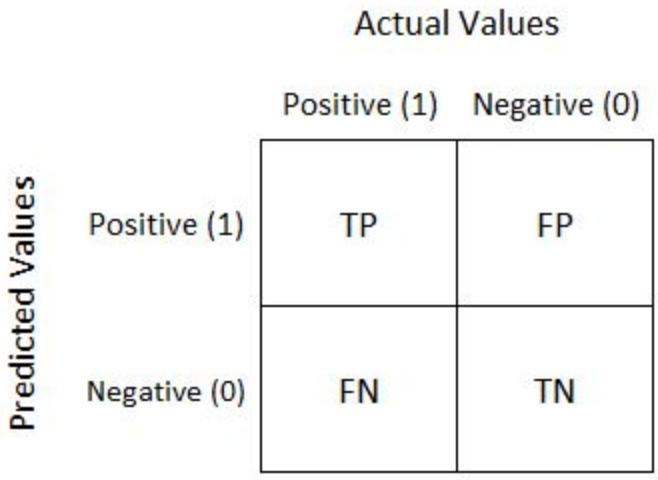
It measures the **impurity or uncertainty present** in the whole data ,this quantity is required fo the information gain formula,lesser the entropy better the perimeter

INFORMATION GAIN:-

It give the information of a **particular root node future give us about the final outcome** .High information gain is better the selection of root node

CONFUSION MATRIX

Well, it is a performance measurement for machine learning classification problems where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.



INFERENTIAL STATISTICS

- **POINT ESTIMATION:-**In statistics, point estimation involves the use of sample data to calculate a single value which is to serve as a "best guess" or "best estimate" of an unknown population parameter. More formally, it is the application of a point estimator to the data to obtain a point estimate

  Estimator:- function that calculator the single value of the sample data
  Estimate:- value of the estimator is estimate
- **MARGINAL INTERVAL(sampling error):-**

  Difference between point estimate and the actual population parameter(real) value is called **sampling error**
- **INTERVAL ESTIMATE:-**

In statistics, interval estimation is the use of sample data to calculate an interval of possible values of an unknown population parameter; this is in contrast to point estimation, which gives a single value.

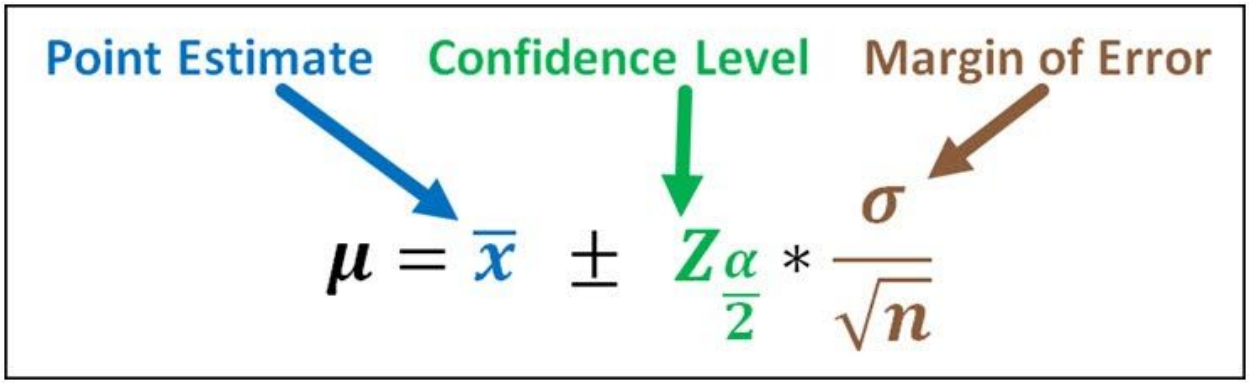
- **CONFIDENCE INTERVAL:-**
  Measure of your confidence that the interval estimate the population mean
- **CONFIDENCE LEVEL:-** how confidence that the value lies in the confidence interval you are at the prediction

MARGIN OF ERROR:-

The **margin of error** is a statistic expressing the amount of random sampling error in the results of a survey. The larger the margin of error, the less confidence one should have that a poll result would reflect the result of a survey of the entire population.

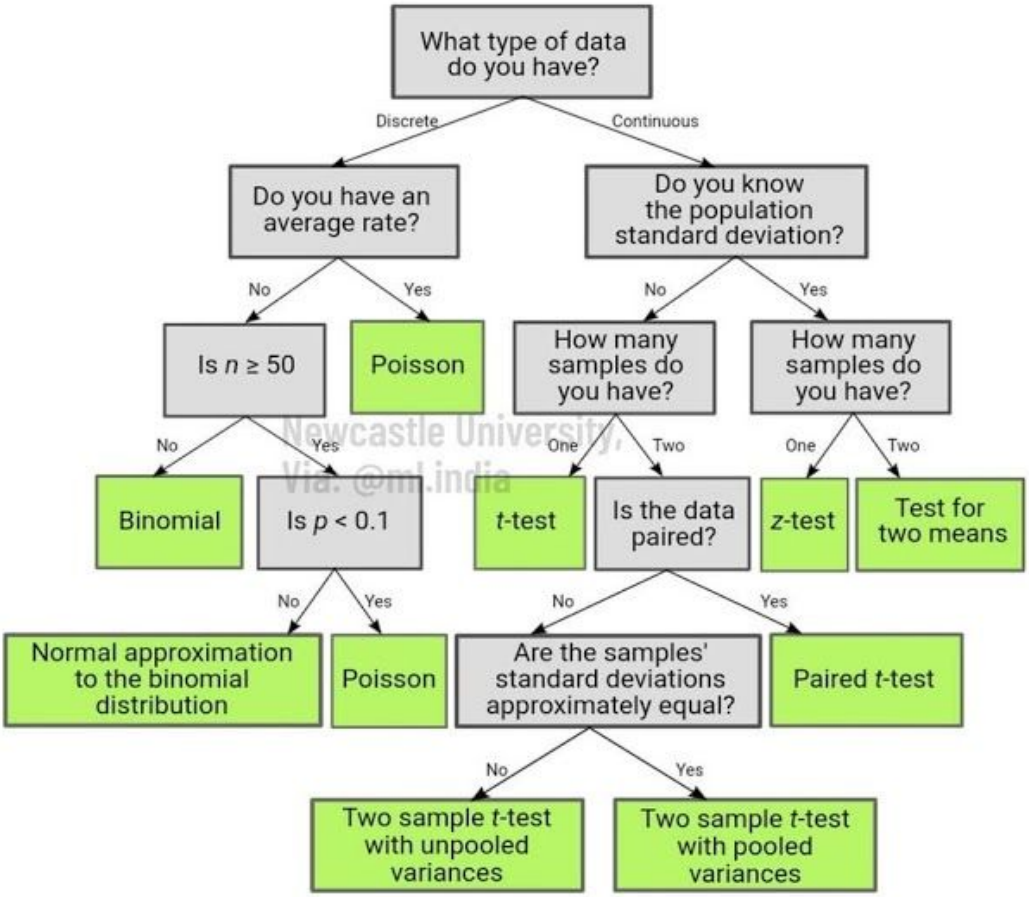
ML &AI

Jul 25, 2020–Jan 26, 2021



HYPOTHESIS TESTING

Hypothesis Testing:

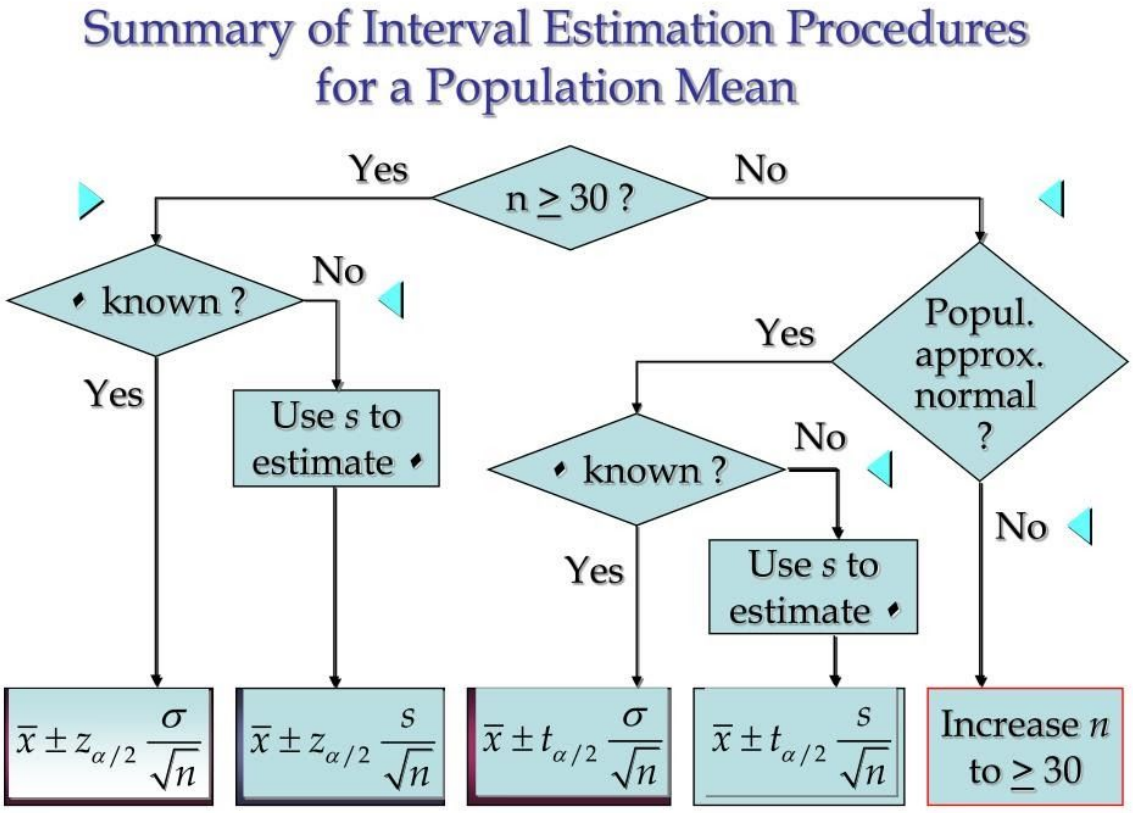


Hypothesis testing to formally **check whether the hypothesis(taken in the previous step) is accepted or rejected**

Steps for hypothesis testing

- State the hypothesis
- Formulate and analysis plan
- Analyse sample data
- Interpret results
- **CRITICAL VALUE or THRESHOLD VALUE:-** set a value of probability for the hypothesis
- **T-TEST**:-A t-test can only be used when comparing the means of two groups (a.k.a. pairwise comparison). If you want to compare more than two groups, or if you want to do multiple pairwise comparisons, use an ANOVA test or a post-hoc test.
- **Z-TEST:-** In a z-test, the sample is assumed to be normally distributed. A z-score is calculated with population parameters such as “population mean” and “population standard deviation” and is used to validate a hypothesis that the sample drawn belongs to the same population

- **Z-SCORE :-** number of standard deviations the data point is far from mean
  critical value gives the vale of z in the sheet for us to know where the data is located compared to standard deviations



- **NULL HYPOTHESIS:-**Taken hypothesis is true similar to universal truth already existing condition
- **ALTERNATE HYPOTHESIS:-**Taken hypothesis is contradicting the take hypothesis

P-VALUE AND STATISTICAL SIGNIFICANCE

The level of statistical significance is often expressed as a *p*-value between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

- A *p*-value less than 0.05 (z or t)(typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

However, this does not mean that there is a 95% probability that the research hypothesis is true. The *p*-value is conditional upon the null hypothesis being true is unrelated to the truth or falsity of the research hypothesis.

BASICS

BALANCED DATASET:-
Number of data points of each class is same or approximately same it is called balanced data set
IMBALANCED DATASET:-
Number of data points of each class is not same or approximately same it is called balanced data set
Univariate analysis:-Analysing every variable one by one
Ex:-PDF,CDF,box-Plot,violin Pot
Bivariate analysis:-Analysing two variables at time
Ex:-scatterplot, pair plot
Multivariate analysis:-more than 2 variable are analyzed
Ex:-Machine learning

MATHEMATICS IN MACHINE LEARNING

LINEAR ALGEBRA

- **Linear equation:- order 1**
- **Vectors**
- **matrix**

Line ,plane, hyperplane
Definition and how to find the distance from a point
Square,rectangle, circle ,ellipse
Cuboid,ellipsoid,sphere
Hyper cuboid,hyper ellipse

CALCULUS
Multivariate calculus
Differentiation
Gradient descent

- Power Pivot
- Power Query
- Power View